1. Lineær regression

Givet datapunkterne:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

kan man opstille en lineær funktion af formen:

$$f(x) = a \cdot x + b$$

hvor hældningen a og skæringspunktet b er givet ved:

$$a = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$$

$$b = \bar{y} - a \cdot \bar{x}$$

hvor $\bar{x} = \frac{1}{n}\sum x_i$ og $\bar{y} = \frac{1}{n}\sum y_i$.

Det vi kalder ”den bedste rette linje” gennem punkterne er den linje som ”mindste kvadraters metode” finder frem til For at finde formlerne for a og b betragter vi residualerne, hvilket er fejlen eller forskellen mellem den realiserede y-værdi(data i tabel) og den estimerede værdi(beregnet fra modellen). Den bedste rette linje, altså modellen, rammer jo kun tilfældigvis ind i nogle af datapunkterne, men slet ikke alle, og kommer således med en ”fejl” for hvert datapunkt. Residualet i hvert punkt er netop denne ”fejl”, fejl som vi jo lever fint med, da vi får den bedste rette linje og med den kan estimere eller ligefrem forudberegne værdier, der jo ikke ligger i tabellen.

Den samlede sum af disse fejl, fejlenes kvadrater, er altså det, der giver anledning til vi kan finde den bedste rette linje, idet vi går efter at gøre ”fejlsum” mindst mulig.

Inden vi går rigtigt i gang med beviset for a og b, starter vi lige med en definition af sumtegnet og et par små hjælpesætninger om summer. Vi gider ikke skrive ”i=1 og til n” ved alle sumtegnene, men det bør man strengt taget gøre hver gang.

4. Summationsegenskaber

Definition: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

Egenskab 1: $\sum (x_i + y_i) = \sum x_i + \sum y_i$

Egenskab 2: $\sum a \cdot x_i = a \cdot \sum x_i$

Egenskab 3: $\sum c = n \cdot c, \quad \text{hvor } c \text{ er konstant}$

---

5. Beviser

Bevis for Egenskab 1: $\sum (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n)$ Ved brug af associativitet og kommutativitet kan man omstrukturere: $= (x_1 + x_2 + \ldots + x_n) + (y_1 + y_2 + \ldots + y_n) = \sum x_i + \sum y_i$

Bevis for Egenskab 2: $\sum a \cdot x_i = ax_1 + ax_2 + \ldots + ax_n = a(x_1 + x_2 + \ldots + x_n) = a \sum x_i$

Bevis for Egenskab 3: $\sum c = c + c + \ldots + c = n \cdot c$

2. Residualer og kvadratsum

Et residual er forskellen mellem en faktisk y-værdi og den estimerede:

$$r_i = y_i - f(x_i) = y_i - (a \cdot x_i + b)$$

Kvadratsummen $K(a, b)$ er summen af de kvadrerede residualer:

$$K(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

---

3. Minimum og differentiering

For at finde det bedste fit minimeres $K(a, b)$. Det gøres ved at differentiere med hensyn til a og b og sætte begge afledte lig nul:

$$\frac{\partial K}{\partial a} = 0, \quad \frac{\partial K}{\partial b} = 0$$

Løsning af ligningerne giver formlerne for a og b som ovenfor.

$$a = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \quad \text{og} \quad b = \bar{y} - a \cdot \bar{x}$$

Disse udtryk er for koefficienterne a og b i den **lineære regressionsmodel**:

$$y = ax + b$$

Lad os bevise dem én ad gangen.

---

1. Bevis for a:

Formålet er at finde hældningen a, der minimerer summen af kvadrerede fejl:

$$S(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

For at finde minimum, differentierer vi S mht. a og b og sætter begge gradienter lig 0 (de normale ligninger):

1. Afledt mht. a:

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0$$

2. Afledt mht. b:

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

Omskriver vi disse:

$$\partial b \qquad i=1$$

Omskriver vi disse:

(1) $\sum x_i y_i - a \sum x_i^2 - b \sum x_i = 0$

(2) $\sum y_i - a \sum x_i - nb = 0$

Brug nu middelværdier:

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i$$

Løs (2) for b:

Så fås:

- $b = \bar{y} - a\bar{x}$

Sæt dette udtryk for b ind i (1) og isoler a:

$$\sum x_i y_i - a\sum x_i^2 - (\bar{y} - a\bar{x})\sum x_i = 0$$

Forkortes og omformes det, får vi:

$$a = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}$$

2. Bevis for $b = \bar{y} - a\bar{x}$:

Dette er allerede udledt ovenfor direkte fra de normale ligninger.

EKSEMPEL PÅ BRUG AF FORMLERNE:

1. Punkterne:

$$(1,2),\ (2,3),\ (3,2),\ (4,5)$$

---

2. Beregn gennemsnit:

$$\bar{x} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$$
$$\bar{y} = \frac{2+3+2+5}{4} = \frac{12}{4} = 3.0$$

---

3. Beregn tæller i formlen for a:

$$\sum x_i(y_i - \bar{y}) = 1(2-3) + 2(3-3) + 3(2-3) + 4(5-3)$$
$$= 1(-1) + 2(0) + 3(-1) + 4(2) = -1 + 0 - 3 + 8 = 4$$

4. Beregn nævner i formlen for a:

$$\sum x_i(x_i - \bar{x}) = 1(1-2.5) + 2(2-2.5) + 3(3-2.5) + 4(4-2.5)$$
$$= 1(-1.5) + 2(-0.5) + 3(0.5) + 4(1.5) = -1.5 - 1.0 + 1.5 + 6.0 = 5.0$$

---

5. Beregn a:

$$a = \frac{4.0}{5.0} = 0.8$$

---

6. Beregn b:

$$b = \bar{y} - a \cdot \bar{x} = 3.0 - 0.8 \cdot 2.5 = 3.0 - 2.0 = 1.0$$

Resultat:

$$a = 0.8, \quad b = 1.0 \Rightarrow \text{Regressionslinje: } y = 0.8x + 1.0$$